Predicting Outcome of ODI Cricket Games

Kevin Desai
Computer Engineering Department
Sardar Patel Institute of Technology
Mumbai, India

Siddhant Doshi
Computer Engineering Department
Sardar Patel Institute of Technology
Mumbai, India

Surekha Dholay
Computer Engineering Department
Sardar Patel Institute of Technology
Mumbai, India

***Abstract*—Predictingthe result ofa cricketgame has always been challenging, as it involves a wide variety of factors.There are numerous natural parameters which are responsible for the outcome of a game. These parametersmake it very difficult to predict the outcome of the matches. Moreover, outcome of these games is not only based on the team's ability and talent of the players, but it is also based on the venue of the match, weather and pitch conditions. Hence all these factors make predicting cricket matches extremely difficult.We tried to predict the outcome of One Day International games in cricket by proposing a numerical model.A number of factors were considered together to reach to a conclusion. Weights were considered for each factor depending on its importance. The conclusion is in the form of win-loss percentage for both the participating teams at a given venue. Using this statistical model we achieved an accuracy of around 70%.**

Keywords—predictive analysis;numerical model; cricket; data scraping;

I. INTRODUCTION

A One Day International (ODI) is a 50 overs limited cricket game which is played between two international teams. All the matches played in the World Cup follow this format.The international status of a team is determined by the International Cricket Council (ICC).The teams which plays test matches havepermanent ODI status. There are 10 such test playing teams. Cricket is, today known as a sport which generates countlessmoney and is immensely popular. Rules governing this game are complex. But, since cricket is the second most watched sport in the world after football, and enjoys a multimillion dollar industry, the need for predicting cricket games is always there.

II. RELATED WORK

A. Sports Predictions

Before the advent of data science, many sports organizations relied only on human expertise. It was believed that domain experts could effectively make use of their knowledge to convert their collected data into useful information. As the collection of data kept on growing, these organizations kept on finding more practical and reasonable approaches to accurately predict the outcome of the game. By finding right ways to turn data into gold, some sports organizations try to secure a competitive edge over their peers. Hence, as data kept on increasing sports prediction using data mining and machine learning gained a lot of popularity.

B. Data Mining in Football, NFL and Baseball

Prediction in cricket games is relatively new and not explored mainly due to the complexities involved in this sport. Also since rules of ODI cricket have undergone major overhaul in recent years, it is difficult to make a consistent model which can last for a long time. Other sports like football, NFL and baseball have lots of research done in predicting their games. We looked at a number of papers which predicted the outcomes of these sports. BabakHamadani [1] used Support Vector Machine and Logistic Regression to predict the outcome of NFL games. Jim Warner [2] also used machine learning to predict the margin of victory in NFL games. With respect to cricket games, VigneshVeppurSankaranarayanan [3] tried to simulate cricket ODI games using team and player statistics. Different data mining techniques used for result prediction in sports were studied [4].

III. RULES OF ODI CRICKET GAME

Over here we provide a basic overview of rules that govern ODI games.

- An ODI match is played between 2 teams of 11 players each.
- The captain of the toss winning side can either choose to bowl or field first.
- The team batting first tries to score as many runs to set a good target to defend. The innings will last until full 50 overs get completed or the team gets all out (10 out of 11 players get out).
- Each bowler can bowl a maximum of 10 overs except in a rain affected match.Therefore, each team must have to chooseat least five bowlers in order to complete the 50 overs.

- In order to win the game, the team batting second tries to score more runs than the target set by the team batting first. On the other hand, the team bowling second tries to defend the score they have set for the opposition to chase.
- The game is declared as tie if the team batting second has scored equal number of runs as the team batting first even though they both have lost different number of wickets at the end of the game.

These are the major rules which can give a brief understanding about ODI cricket games. There are many more rules and details, but those are not essential in predicting the outcome of the games.

IV. DATA COLLECTION

ESPN has a division of its own dedicated to cricket statistics, known as the espncricinfo.com. We wrote a custom script in PHP and scraped data for 10 years, beginning from 2001 up to first quarter of 2011. The data that we scrapped and imported in MySQL database was not structured properly and hence we had to clean the database so that it suits our model. We carried the following steps to structure the data properly.

- The data we scraped was present in different tables like, grounds, matches and teams. We wrote SQL queries and created one single table known as cricket in which there were different columns like, date of the match, team 1, team 2, winning team, the ground at which the match was played, country in which the stadium is located, the team which won the toss and whether the winning team was playing at home or away.
- During the phase of 10 years there were a lot of teams other than the 10 test playing nations which had played ODI games. Now these teams have played far less number of matches as compared to the 10 test playing nations. These few matches act as outliers in the dataset and often are found to reduce the accuracy of the prediction. Hence we removed all the matches in which even one of the two playing teams was other than the 10 test playing nations.
- In cricket, weather and light conditions are very important. Hence a number of matches are either delayed (till the conditions are perfect again) or abandoned altogether. Abandoned matches also affectaccuracy of the prediction and hence matches in which no result was achieved were removed from the dataset.
- In cases of both the teams finishing by scoring equal number of runs in the stipulated overs, the result of the match is declared as a tie. Since tied matches have not achieved any result in terms of a win or a loss, such matches cannot be taken into account for. Therefore tied matches were also removed from the dataset.

We achieved all of the above steps by writing simple SQL queries.

V. FACTOR SELECTION

Cricket being a complex sport to make accurate predictions, selection of relevant features which make a big difference to the prediction of the outcome of a particular game is extremely important. Let us assume that team A and team B are playing against each other where A is the home team and B is the away team. Based on years of watching and understanding cricket we chose two sets of 12factors (one set for team A and other set for team B) which we think will help to accurately determine the outcome of a particular game. These factors are either expressed in terms of win to total matches or loss to total matches ratio.

Set of factors for team A = {a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}, a_{11}, a_{12}}

Set of factors for team B = {b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9, b_{10}, b_{11}, b_{12}}

1. The simplest of all features is the performance of both the playing teams in past 10 years. (Data we had was of 10 years). Hence performance of the home team and away team in past 10 years makes up for 4 factors.

   a_1 - team A’s win to total matches ratio in past 10 years

   b_1 - team B’s win to total matches ratio in past 10 years

   a_7 - team B’s loss to total matches ratio in past 10 years

   b_7 - team A’s loss to total matches ratio in past 10 years

2. Cricket teams undergo a lot of changes with respect to the players in the team. Some teams change players from match to match. Hence the performance of a team depends on the players in the team. And hence recent form of a team becomes a very important factor. Hence recent forms of the home and away team are 4 more factors we have considered.

   a_2 - team A’s win to total matches ratio in past 2 years (recent form)

   b_2 - team B’s win to total matches ratio in past 2 years (recent form)

   a_8 - team B’s loss to total matches ratio in past 2 years (recent form)

   b_8 - team A’s loss to total matches ratio in past 2 years (recent form)

3. Home advantage is an essential advantage in not just cricket but every other sport. Hence the

performance of the home team in last 10 years and also the recent performance make up for 2 more features. On the other hand, the away conditions might prove to be disadvantageous for some teams. Hence we consider 2 more factors for away conditions.

a_3 - team A's win to total matches ratio in past 10 years in home conditions

a_4 - team A's win to total matches ratio in past 2 years in home conditions (recent form)

b_3 - team B's win to total matches ratio in past 10 years in away conditions

b_4 - team B's win to total matches ratio in past 2 years in away conditions (recent form)

4. It happens quite sometimes that a certain team, if it plays at a particular stadium is slated to win no matter what. Therefore the stadium or the ground at which the match is played also should be considered. Hence the home team's performance at the given stadium for recent and past performance is also taken into account. This gives us following 4 factors:

   a_5 - team A's win to total matches ratio at given stadium in past 10 years

   a_6 - team A's win to total matches ratio at given stadium in past 2 years (recent form)

   b_5 - team B's win to total matches ratio at given stadium in past 10 years

   b_6 - team B's win to total matches ratio at given stadium in past 2 years (recent form)

5. Cricket has far less number of teams as compared to other sports like football or NFL. Hence the same teams play against each other a lot of times. It is essential to take into account all of the past matches between the two teams in recent and past irrespective of where they have played. For example, if Team A loses against Team B every time but wins against all the other teams, the probability of Team A losing to Team B is higher because it always loses against Team B.

   a_9 - team A's win to total matches ratio against team B in past 10 years

   a_{10} - team A's win to total matches ratio against team B in past 2 years (recent form)

   b_9- team B's win to total matches ratio against team A in past 10 years

   b_{10} - team B's win to total matches ratio against team A in past 2 years (recent form)

6. Also the recent and the past encounters between the two contesting teams in the home country have to be considered. And therefore, for example, Team A's performance against Team B in Team B' country is taken into consideration.

   a_{11} - team A's win to total matches ratio against team B in team A's home condition in past 10 years

   a_{12} - team A's win to total matches ratio against team B in team A's homecondition in past 2 years (recent form)

   b_{11} - team A's loss to total matches ratio against team B in team A's home condition in past 10 years

   b_{12} - team A's loss to total matches ratio against team B in team A's home condition in past 2 years (recent form)

For neutral games (games in which neither of the two teams are playing in their home countries) a few features change. Instead of the home country, performances of the teams in the country they are playing are taken into consideration. The other features remain, more or less the same.

VI. METHODOLOGY

As explained above we considered a total of 12 factors for each team for predicting the outcome of a game. For each of the factors, based on its importance we assigned them weights. For example, the importance of recent matches is more than that of matches of past 10 years, and hence factors in which recent matches are considered have been assigned greater weights. Also, it has been observed that the home and away conditions greatly influence the outcome of the game. Hence the factors related to home and away conditions have been assigned greater weights.

Set of weights = $\{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9, w_{10}, w_{11}, w_{12}\}$

Here each $w_i$is associated with both $a_i$and $b_{i.}$

Initially random values were assigned to weights. These values were adjusted and the code was then run until we reached the maximum accuracy. We plan to use machine learning or artificial neural networks to adjust the weights in future.

We have proposed the following numerical model :

$$teamA = \sum_{i=1}^{12} (a_i * w_i)$$

$$teamB = \sum_{i=1}^{12} (b_i * w_i)$$

$$\%teamA = teamA * 100 / (teamA + teamB)$$

$$\%teamB = teamB * 100 / (teamA + teamB)$$

Where *%teamA* is the winning percentage of team A and *%teamB* is the winning percentage of team B for the away team.

We first multiply each member of set of weight to each corresponding member of set of factors for team A to find the value of *teamA*. Similarly, we find the value of *teamB*. Then we calculate the win percentage of teamA and teamB.

Let us take one example to understand the working of this numerical model. Let us assume that India is playing against Australia at Wankhede (a stadium in India). Here team A is India and team B is Australia.

Let us assume that India has won 140 out of 320 matches whereas Australia has won 200 out of 300 matches in past 10 years.

Let $w_1 = 0.5, w_7 = 0.5$

So,

$a_1 = 140 /320$

$b_1 = 200/300$

We have not taken tied and abandoned matches into account. So,

$a_7 = 100 /300$

$b_7 = 180/320$

If we consider only the above factors we can say that,

$teamA = a_1 * w_1 + a_7 * w_7 = 0.3854$

$teamB = b_1 * w_1 + b_7 * w_7 = 0.6146$

$teamA < teamB$

$\%teamA < \%teamB$

Now suppose if Australia has been consistently losing in past years whereas India's performance has been excellent in past 2 years. So it would not be right to say that Australia will win based on past 10 years performance. That is why, we have considered recent form in our model. Let us assume that India has won 52 matches out of 80 matches whereas Australia has won 28 matches out of 60 in last 2 years.

$a_2 = 52 /80$

$b_2 = 28/60$

We have not taken tied and abandoned matches into account. So,

$a_8 = 32 /60$

$b_8 = 28/80$

$a_2 > b_2$

$a_8 > b_8$

Since this model plans to give more weightage to recent form

Let $w_2 = 2.0, w_8 = 2.0$

$X = a_2 * w_2 + a_8 * w_8 = 2.3667$

$Y = b_2 * w_2 + b_8 * w_8 = 1.6333$

So now if we consider these factors,

$teamA(new) = teamA + X = 2.7521$

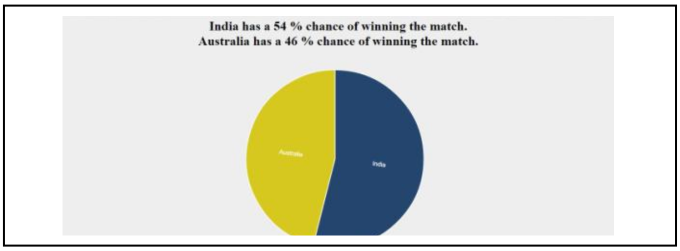
$teamB(new) = teamB + Y = 2.2479$

So now,

$teamA > teamB$

$\%teamA > \%teamB$

This model will take into consideration all other different factors in similar way which would give us the final value of *teamA* and *teamB*and help us to calculate their winning percentages.



We developed a simple UI to access the code written in PHP. The user was given options to select both the teams and the stadium at which the match is going to take place. On clickingthe submit button, the values were submitted to the PHP script which calculated the results based on the factors mentioned above and returned the results in a form of visualization, which is displayed above.

VII. RESULTS

Our dataset has matches up till 3rd March 2011, after which we predicted 30 games involving test playing nations.

Out of the thirty games we predicted 21 turned out to be correct and 9 matches were predicted wrongly.

That is a prediction accuracy of exactly 70%.

Predicting ODI matches		
Total Matches Predicted	***Correct***	***Wrong***
30	21	9

VIII. FUTURE SCOPE

Our numerical model does not take into account numerous factors which can improve the accuracy of the model even more.

- Some teams might be good in chasing the target whereas some teams might be good in defending the target. Also, the pitch conditions can change over the course of the game. Hence, the toss factor can play important role in deciding the outcome of cricket games. Hence in future we plan to consider the toss factor also.

- In this model we just predict the win loss percentages of both the teams. But in future we plan to predict the number of runs scored and even the margin of victory if possible.
- Also, we can improve this model much more by using different algorithms like Support Vector Machines and Logistic Regression.
- The model can be improved for better prediction of matches at a neutral venue.
- To calculate assigned weights programmatically and use Artificial Neural Networks to iterate over and recalculate weights according to the dataset.

REFERENCES

[1] Babak Hamadani, ‘Predicting the outcome of NFL games using Machine Learning’,http://cs229.stanford.edu/proj2006/BabakHamadani-PredictingNFLGames.pdf.

[2] Jim Warner: ‘Predicting Margin of Victory in NFL games’, http://www.cs.cornell.edu/courses/cs6780/2010fa/projects/warner_cs6780.pdf.

[3] Vignesh Veppur Sankaranarayanan, Junaed Sattar and Laks V. S. Lakshmanan ‘Auto-play: A Data Mining Approach toODI Cricket Simulation and Prediction’

[4] http://www.academia.edu/5288042/A_Review_of_Data_Mining_Techniques_for_Result_Prediction_in_Sports